Clustering & PCA Assignment
Part II

Question 1: Assignment Summary

Problem Statement:

- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. After the recent funding programs, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively to fund the countries that are in need of aid.

Solution Methodology:

- There are nine attributes that needs be analyzed to determine which countries need the aid , since the variables are more it's difficult to analyze which variable/variables contribute most to the analysis.
- PCA will be used to find the linear combination between the attributes,to remove multicollinear data, to find the Principal Components which help in Dimensionality Reduction without dropping any attributes.
- To determine the Principal Components required for the analysis **Scree Plot** is used.
  So from this plot it is evident that **"3 PC"** are enough to explain about 90% of variance in the data.
- Before performing clustering **Hopkins Score** was calculated to check how good the data set is eligible for clustering. A Hopkins Score of variable 70 to 75% was obtained implying data was just good for clustering.
- K-Means Clusters was done on 3 PC's. In K-Means Clustering to determine the number of clusters Silhouette Score and Elbow Curve is used. From Silhouette Score a cluster of 3 being maximum value was obtained. From Elbow Curve the curve was seen at 4 and 5.
  Cluster with 3 doesn't make sense, cluster with 4 and 5 was considered. Cluster with k=4 and k=5 produced almost same results. So 4 clusters were considered. Cluster 0 and cluster 3 had higher child mortality rates, low income and low GDP per capita. There by telling these countries are in highest need of aid. Cluster 1 and Cluster 2 have high child mortality, high income and high GDP per capita.
- Hierarchical Clustering was also performed on the 3 PC's. From Dendogram the tree cut with 4 cluster seemed most reliable with low GDP per capita, low income and high child mortality were grouped to Cluster 0 there by making cluster 0 in need of

aid. Cluster 2,3 have low child mortality and high income, gdpp is also relatively high. Cluster 1 has high to low wide spread of child mortality and low income and low gdpp.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

K-Means Clustering	Hierarchical Clustering
➢ Division of objects into clusters such that one object belongs to only one cluster	➢ Tree like structure or parent-child relationship.
➢ Starts with random choice of cluster centre, yields different clusters on different run.	➢ Consistent result as there is no concept of cluster centers/centroids.
➢ Requires prior knowledge of k.	➢ No prior knowledge is required we can cut the tree at any point as per requirement.
➢ Time complexity is 0(n)	➢ Time complexity is O(n2)
➢ Large datasets.	➢ Small datasets.
➢ Each time tweaks are necessary to be able to build fair model.	➢ Not necessary just run the model without modification.
➢ There is just iterations and moving centroids no model building is happening.	➢ Requires good RAM as model building is complex.

b) Briefly explain the steps of the K-means clustering algorithm.

- First we need to choose the K random points which are new cluster centers for the data set.
- Assign the data points to the nearest cluster center. Euclidean Distance is used for the purpose.
- For each cluster , calculate the new cluster centers which will be the mean of all the cluster members.
- Assign the data points to the new cluster centers obtained.
- Keep iterating through the steps 3 & 4 until no changes in the cluster centers can be made. At this stage we arrive at the optimal cluster points.
- Graphical representation of K-Means Clustering.

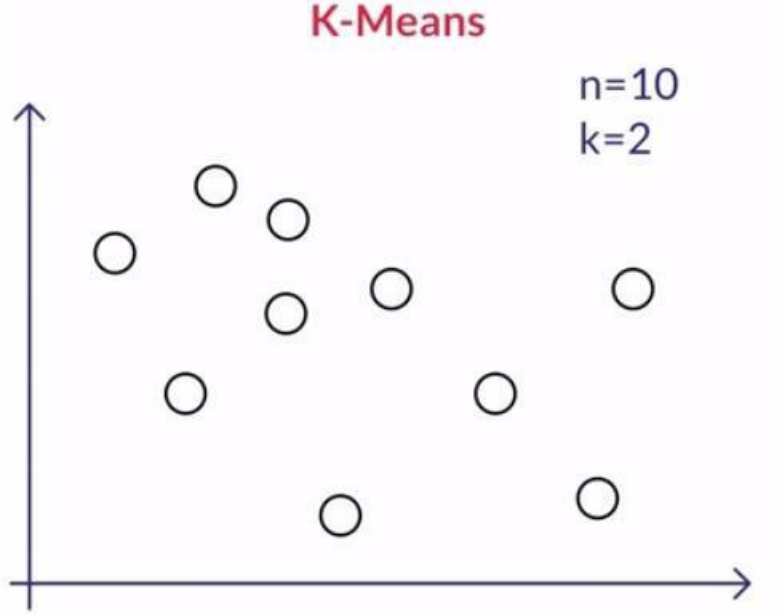


Fig 2: A set of 10 points to be divided into 2 clusters

We begin with choosing 2 random points as the 2 cluster centers.

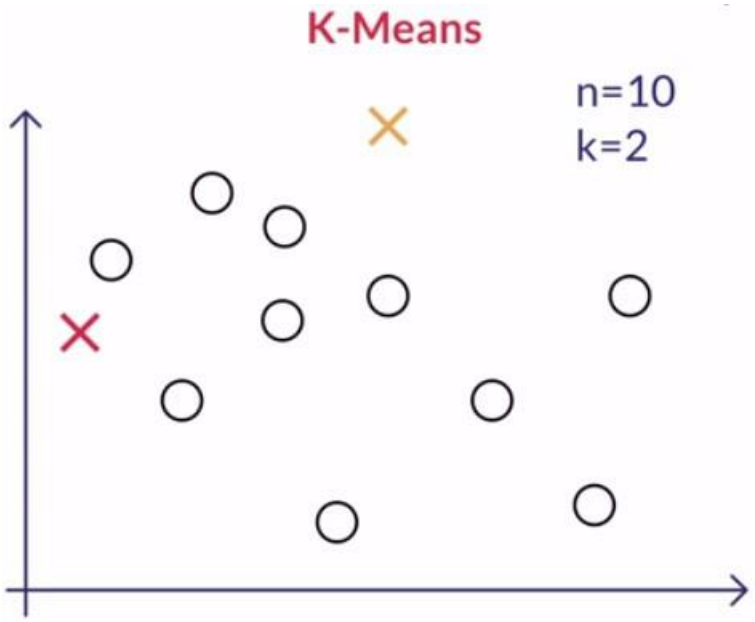


Fig 3: Choosing K random initial cluster centers

We then assign each of the data points to their nearest cluster centers based on the Euclidean distance. This way all the points are divided among the K clusters.

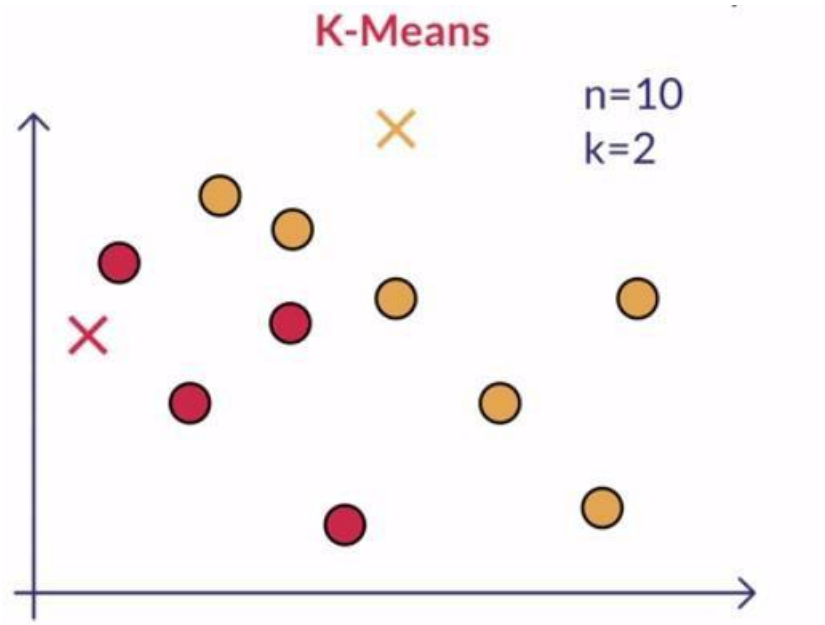


Fig 4: Assigning each data point to their nearest cluster centre

Now we update the position of each of the cluster centers to reflect the mean of each cluster.

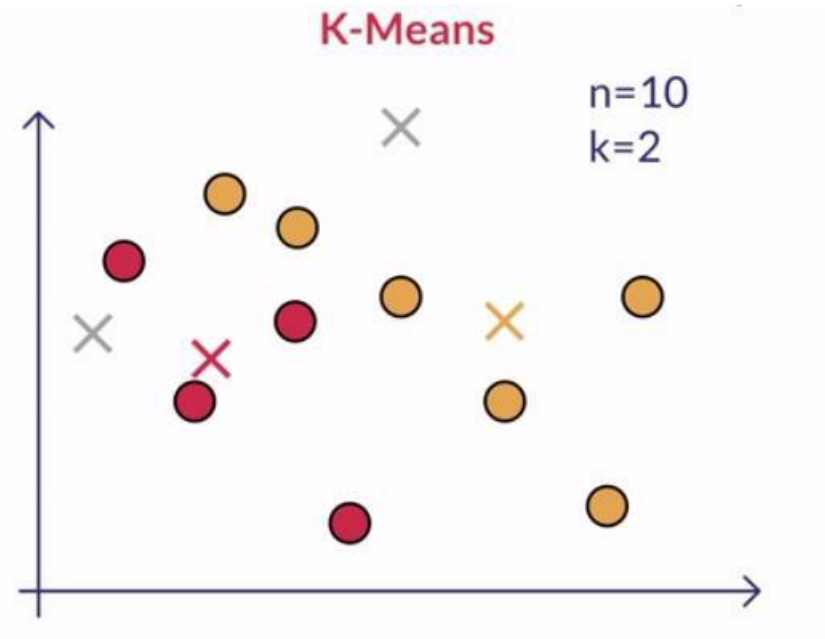
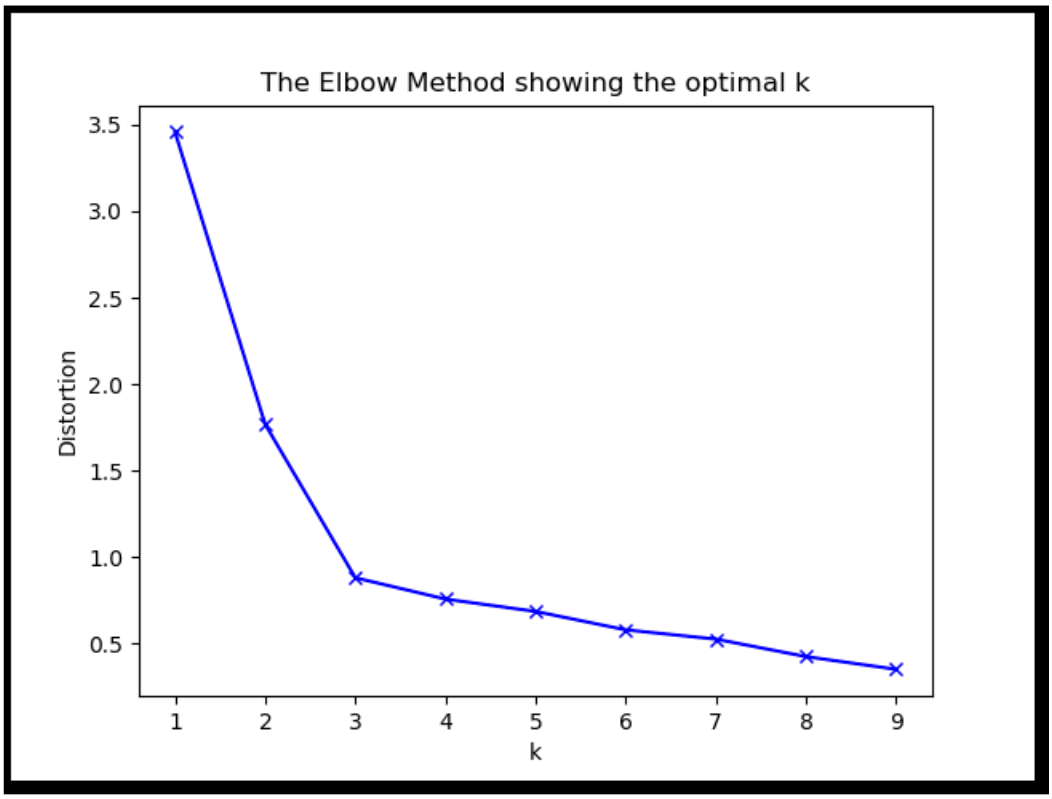


Fig 5: Updating the cluster centers

This process continues iteratively till the clusters converge; that is, there are no more changes possible in the position of the cluster centers. At this point, we achieve the two optimal clusters.

c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

- In K-means there are two statistical methods of choosing the K value and are:
    1. **Silhouette Score** : For each value of K, we calculate the average Silhouette Score And plot the graph of average Silhouette Score against the value of K. The peak value in the curve determines the optimal number of clusters.
    2. **Elbow Curve**: For each value of K, we calculate the total within-cluster sum of Squares. Plot the graph of within-cluster sum of squares against The value of K. The elbow represents the optimal value of K.



From the above figure the optimal value of K is 3.

- As per the business aspect we can enquire the client Manager to give idea of how many categories is he looking for. As he/she will be the best person to give insight in deciding the optimal clusters that he/she is looking for.

d) Explain the necessity for scaling/standardisation before performing Clustering.

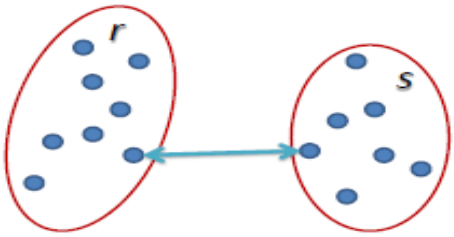
- Incomparable units. Height vs weight. You cannot compare, so the default decision is to standardize (equalize variances); it is "default" on the grounds of thought parsimony: "every unique aspect of nature is assumed to have same, unit variability of observations".
- Same units, irrelative features. Height vs circumference. These are clearly independent (conceptually, not statistically) phenomena of reality. Their same-unitness seems a coincidence. It would be silly to compare between the two values. The default decision is to standardize the features.

- Same units, juxtaposed features. Length of right arm vs of left arm. We could naturally compare the two lengths if we need so, they two are interchangeable, in a sense. The default decision is to leave variances as is (no matter how much they differ). Because "leave nature under study be how it is".
- Undecided whether 2 or 3. Length of arm vs length of leg. We could compare these but we are not interested in that, rather, we prefer to see the lengths as separate dimensions (albeit not irrelative phenomena). Feature-conceptual decision (whether standardize or leave) is impossible. Other, method-driven or goal-driven or criterial-driven1 considerations would dictate the choice in a concrete situation. No default solution and the decision could be difficult to make. Some considerations might resolve the problem by providing an insight that the case is actually 2 or 3.

e) Explain the different linkages used in Hierarchical Clustering

Single Linkage:

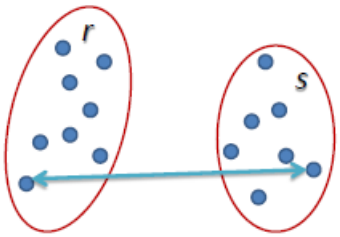
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkage

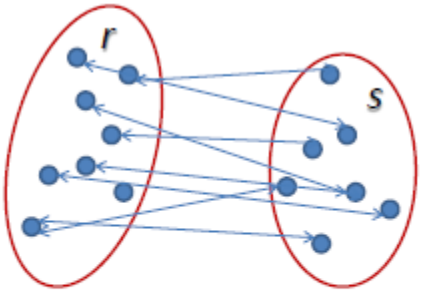
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

- Dimensionality Reduction: PCA basically helps in Feature elimination and Feature Selection without dropping the variables. If there are 100 features on which the analysis has to be carried out PCA calculates the Principal Components which are linear combinations, non-multicollinear , orthogonal, uncorrelated and maximum variance data. We need not have to choose all the Principal Components, select the PC's that can explain maximum variance in the data by taking Cumulative sum of PC's.
- For creating uncorrelated feature that can be input to a predictor model: with smaller number of uncorrelated features, the modeling is faster and more stable.

- For Data Visualization and EDA : If there are lots of variables in the data to visualize and explore.

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Basis transformation:

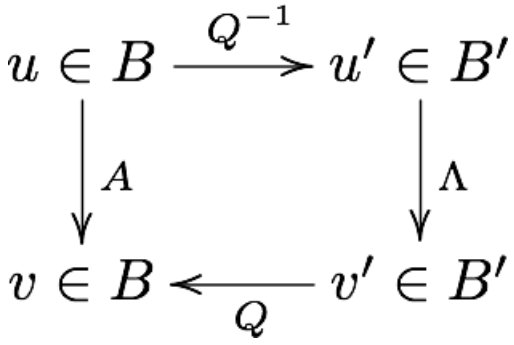
We can transform the original data set so that the eigenvectors are the basis vectors and find the new coordinates of the data points with respect to this new basis

This is the change of basis transformation covered in the Linear Algebra module. First, note that the covariance matrix is a real symmetric matrix, and so the eigenvector matrix is an orthogonal matrix.

Linear algebra review for change of basis

Let's consider two different sets of basis vectors B and B′ for R2. Suppose the basis vectors for B are u,v and that the basis vectors for B′ are u′,v′. Suppose also that the basis vectors u′,v′ for B′ have coordinates u′=(a,b) and v′=(c,d) with respect to B. That is, u′=au+bv and v′=cu+dv since that's what vector coordinates mean.Suppose we want to find out what the coordinates of a vector w=(x′,y′) in the B′ basis would be in B. We do some algebra:So expressing in matrix form.Since [w]B′=(x′,y′), we see that the linear transform we need to change a vector in B′ to one in B, we simply multiply by the change of coordinates matrix P that is the formed by using the basis vectors as column vectors, i.e.To get from B to B′, we multiply by P−1.To convert from the standard basis (B) to the basis given by the eigenvectors (B′), we multiply by the inverse of the eigenvector matrix V−1. Since the eigenvector matrix V is orthogonal, VT=V−1. Given a matrix M whose columns are the new basis vectors, the new coordinates for a vector x are given by M−1x. So to change the basis to use the eigenvector matrix (i.e. find the coordinates of the vector x with respect to the space spanned by the eigenvectors), we just need to multiply V−1=VT with x.

Graphical illustration of change of basis:



$A = Q^{-1}\Lambda Q$

Suppose we have a vector u in the standard basis B , and a matrix A that maps u to v, also in B. We can use the eigen values of A to form a new basis B′. As explained above, to bring a vector u from B-space to a vector u′ in B′-space, we multiply it by Q−1, the inverse of the matrix having the eigenvectors as column vectors. Now, in the eigenvector basis, the equivalent operation to A is the diagonal matrix Λ - this takes u′ to v′. Finally, we convert v′ back to a vector v in the standard basis by multiplying with Q. Principal Components are simply the eigenvectors of the covariance matrix used as basis vectors. Each of the original data points is expressed as a linear combination of the principal components, giving rise to a new set of coordinates.
For example, if we only use the first column of ys, we will have the projection of the data onto the first principal component, capturing the majoriyt of the variance in the data with a single featrue that is a linear combination of the original features.
We may need to transform the (reduced) data set to the original feature coordinates for interpreation. This is simply another linear transform (matrix multiplication).

Variance as Information

TL;DR

The total variance is the sum of variances of all individual principal components.The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance. For several principal components, add up their variances and divide by the total variance. Consider the example below:
In case of PCA, "variance" means summative variance or multivariate variability or overall variability or total variability. Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

1.343730519	-.160152268	.186470243
-.160152268	.619205620	-.126684273
.186470243	-.126684273	1.485549631

Now, PCA replaces original variables with new variables, called principal components, which are orthogonal (i.e. they have zero covariations) and have

variances (called eigenvalues) in decreasing order. So, the covariance matrix between the principal components extracted from the above data is this:

1.651354285	.000000000	.000000000
.000000000	1.220288343	.000000000
.000000000	.000000000	.576843142

Note that the diagonal sum is still 3.448, which says that all 3 components account for all the multivariate variability. The 1st principal component accounts for or "explains" 1.651/3.448 = 47.9% of the overall variability; the 2nd one explains 1.220/3.448 = 35.4% of it; the 3rd one explains .577/3.448 = 16.7% of it. So, what do they mean when they say that "PCA maximizes variance" or "PCA explains maximal variance"? That is not, of course, that it finds the largest variance among three values 1.343730519 .619205620 1.485549631, no. PCA finds, in the data space, the dimension (direction) with the largest variance out of the overall variance 1.343730519+.619205620+1.485549631 = 3.448. That largest variance would be 1.651354285. Then it finds the dimension of the second largest variance, orthogonal to the first one, out of the remaining 3.448-1.651354285 overall variance. That 2nd dimension would be 1.220288343 variance. And so on. The last remaining dimension is .576843142 variance. Mathematically, PCA is performed via linear algebra functions called eigen-decomposition or svd-decomposition. These functions will return you all the eigenvalues 1.651354285 1.220288343 .576843142 and corresponding eigenvectors at once.

c) State at least three shortcomings of using Principal Component Analysis.

- Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.
- Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components. For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be

applied.PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScaler from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

- Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.